DOCUMENT RESUME

ED 402 779 FL 024 339

AUTHOR Zavrel, Jakub; Veenstra, Jorn
TITLE The Language Environment and Syntactic Word-Class Acquisition.
PUB DATE 96
NOTE 11p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classification; Discourse Analysis; Foreign Countries; *Language Acquisition; *Language Patterns; Language Research; *Linguistic Theory; Newspapers; News Writing; *Syntax; *Vocabulary
IDENTIFIERS *Semantic Bootstrapping Hypothesis (Pinker); Wall Street Journal

ABSTRACT

A study analyzed the distribution of words in a three-million-word corpus of text from the "Wall Street Journal," in order to test a theory of the acquisition of word categories. The theory, an alternative to the semantic bootstrapping hypothesis, proposes that the child exploits multiple sources of cues (distributional, semantic, or other) that are readily available in an information-rich environment, to acquire syntactic categories of words. This study investigates the quantity of information that is available in a specific language environment. The Lexical Space approach to categorization, which looks at the position of each word in lexical space and the structure of the resulting information, was used. Results indicate that co-occurrence and adjacency of words is highly indicative of their syntactic category, and that the high number of possibly contradictory context features poses no serious problem for bootstrapping. Contains 15 references. (MSE)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Jakub Zavrel
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 402 779

# The Language Environment and Syntactic Word-Class Acquisition

**Jakub Zavrel, Tilburg University**
**Jorn Veenstra, Tilburg University**

In this paper we discuss a distributional analysis of the Wall Street Journal corpus. It is shown that the distributional properties of words in this corpus form a rich source of information for the derivation of syntactic word-classes. If there is so much information in the language environment of children, it would be unwise for them (and for us) to ignore this source of information.

## 1. Introduction

Syntactic categories of words in the lexicon are the building blocks from which any type of grammar must be built, and knowledge of these building blocks is an essential part of human language processing. An important research question in Cognitive Science is the question about the nature and origins of this knowledge. So far, a mostly nativist stance on this issue has been taken. The categories are assumed to be prewired in the child's cognitive system at birth. Since this cannot be a full explanation of how children learn to master the syntactic categories in their particular language, this assumption has the consequence that we must shift our attention to the process of language acquisition, regardless of whether we agree with this nativist assumption or not. Namely, if the syntactic categories are indeed innate, we must go on to answer the question of which steps the child must take in order to hook up its internal representations to the actual language it is exposed to. If they are not innate, we must explain how it is possible that they emerge by exposure to language.

In this paper we will take as a starting point one of the most explicit accounts of category acquisition, a theory with a nativist flavor proposed by Pinker (1984). In discussing his proposal, we will try to convince the reader that Pinker's *Semantic Bootstrapping Hypothesis* is by no means the only possible solution to the acquisition problem. To support this claim, we will argue that a form of distributional bootstrapping, as put forward by e.g. Maratsos & Chalkley (1981), might also be viable contrary to Pinker's claim. For this purpose we will present data about the distributional analysis of a large corpus of text with the help of a method which we call the *Lexical Space*. We will propose an alternative theory of category acquisition based on the concept of Lexical Space. According to this theory the child exploits multiple sources of cues—be they distributional, semantic or other—that are readily available to it in an information-rich environment, to acquire syntactic categories of words. Our research investigates how much information the child has available in this language environment. If there is indeed enough information ready at hand to bootstrap into a primitive classification of syntactic categories it can also be questioned whether children need a prewired representational system for syntactic categories.

### 1.1 Bootstrapping of syntactic categories

Syntactic categories are commonly defined by stating that *two words share the same syntactic category, iff they can be substituted for each other freely, i.e. without causing a violation of syntactic well-formedness*. However, this definition is of little use to the child, because there can be an indefinite amount of possible contexts, and the child cannot make the proper abstractions to characterize the contexts, precisely because it has not acquired syntax yet. Thus we are faced with a *bootstrapping problem*. That is, if the child could get its system started in some syntax-independent way at least a little bit, it could then use this primitive syntax to learn the adult syntax by what Pinker (1984) calls "Stucture Dependent Distributional Learning" (SDDL).

FL024339

One possible solution for the child lies in the fact that syntactic categories correspond to semantic distinctions. The pairing of nouns and objects, verbs and actions, and adjectives and properties should enable the child to make the right hypotheses about the categories of the words. Maratsos & Chalkley (1981) and Maratsos (1982) criticized this view, noting that the semantics of word-classes are not always systematic. There are 'action nouns' such as "flight" or "kiss", 'non-action verbs' such as "like" and "know" and 'action adjectives' such as "noisy". According to Maratsos and Chalkley such inconsistencies might lead the child's hypotheses astray. As an alternative they make a case for *distributional bootstrapping*. They observe that in fact the members of a word-class are the same in that they can occur in the same slots in simple morpho-syntactic patterns. As an example: all verbs can occur with the ending -ed and -ing, and they can mostly co-occur with auxiliary verbs. Every word-class has its own prototypical distributional behavior.

Maratsos (1982 p.258) thinks that children use this distributional analysis in combination with semantics to bootstrap into syntactic word-classes:

> This scheme thus builds grammatical categories on the basis of overall sequential similarity, but also on the basis of semantic function and analysis of the parts of the sequence. It is hence not a purely distributional analysis.

Pinker (1984) has replied to this proposal by presenting a model of bootstrapping through a semantic route. He questions the validity of the distributional bootstrapping hypothesis. A point to which we shall return below. He credits the child with an innate knowledge of syntactic categories, together with knowledge about the semantic core of each category. Because the child acquires the meaning of words, and therefore knows which words are objects, which are actions and which are properties it can make a tentative assignment of words to syntactic categories. According to this *semantic bootstrapping hypothesis* a child acquires the syntactic categories of words via semantics; object-words map to nouns, actions-words to verbs, property-words to adjectives, etcetera. (A closely related variant of the semantic bootstrapping hypothesis is that of Braine (1992), who does not assume the innateness of syntactic categories, but only of semantic categories and induces the syntactic categories from their semantic basis.)

Pinker defends the semantic bootstrapping hypothesis against the arguments given by Maratsos. He does not regard the fact that e.g. not all action-words are verbs, as a serious problem, as long as a majority of them are. The other case, viz. that some verbs do not refer to actions will initially excempt them from the verb category of the child, but this will be fixed as soon as SDDL gets underway. At the same time Pinker criticizes distributional bootstrapping on a number of issues. He states that the distributional approach suffers from a learnability problem and an efficiency problem. The first roughly means that there is not enough information in the distribution of words to bootstrap syntactic categories. As Pinker (1984 p.55) states it:

> . . . the properties that the child can detect in the input—such as the serial positions and adjacency and co-occurrence relations among words—are in general linguistically irrelevant.

His second objection concerns the fact that the child does not know beforehand which features it should pay attention to. To try all combinations of a huge space of possible features would be computationally extremely inefficient. Moreover, Pinker expects that many cues will contradict each other.

In a later paper Maratsos (1990) argues that children cannot rely on distributional information to bootstrap into the category of nouns, because there are three different subclasses of nouns: Common nouns, Pronouns and Proper nouns. These three subclasses have different

distributional properties. Hence the class of nouns cannot be bootstrapped from distributional properties of the words. Maratsos concludes that children can use distributional bootstrapping for all word-classes, except the nouns.

### 1.2 Distributional analysis

In the above we have given some objections to the distributional bootstrapping of syntactic categories, as given by Pinker and Maratsos.

1. There is not enough information in the distribution of words. The information that is available is irrelevant (Pinker 1984).

2. There are too many (possibly contradictory) features available for the child. Which information should it attend to (Pinker 1984)?

3. Some categories (such as nouns) consist of a number of subclasses with different distributions (Maratsos 1990).

Note that these objections to distributional bootstrapping focus on the information available to the child rather than the mechanisms used by the child. In our approach to the acquisition of syntactic categories we look at the information available in the environment of the child, and we want to argue that the child can use any source of information—semantic, distributional and e.g. prosodic, phonetic, morphological and situational information, if it contains powerful cues that can be useful for language acquisition. A logical possibility is that the particular blend of information sources used can even differ from language to language, much as in the cue-competition model of Bates & MacWhinney (1987). If there is enough information in the distribution of words to—at least partially—bootstrap into the syntactic categories of words, it would be unwise for children to discard this information. So, to settle the status of distributional information we propose to examine the language environment of a child by analysing the distributional information in a large corpus of natural language using the method of the *Lexical Space*.

## 2. Testing the distributional environment

In the following sections we will discuss our experiments with the Lexical Space. These experiments replicate to some extent the evidence found in other recent work in Computational Linguistics (e.g. Finch (1993), Hughes (1994), Schütze (1995), Redington, Chater & Finch (1995)) which shows that distributional information from large text corpora is very useful as a source of cues for syntactic categories (at least in English). The method we present allows us, in contrast to the works cited, to relate syntactic categories directly to the structure implicit in the spatial relations between word distributions, without constructing an indirect measure of similarity first, such as a hierarchical cluster tree, and relate that to the syntactic categories. This is based on the view that even syntactic categories do not necessarily have to be mutually exclusive equivalence classes, but that they can exhibit a more gradual or continuous structure as well, and that a discrete clustering is likely to introduce additional inaccuracies.

With respect to the careful interpretation of our results it should be noted that the Lexical Space is not intended to be a model of the child's actual acquisition mechanisms, but only a method for the examination of the information available in the language environment. However, we will draw some speculative parallels in the final section.

### 2.1 The Lexical Space

Similar to the approach of Schütze (1995) we represent words as points in a highdimensional space. The position of a word is determined by its observed distribution. In order to be relevant to the bootstrapping discussion, the distributions that are taken into account must be such that they do not presuppose any knowledge of syntax. In the experiments presented here we have focused on the directly neighboring words as a context. Ironically, this is exactly the kind of distributional information that Pinker views as "linguistically irrelevant". For each focus-word we count the number of times that words from a set of context-words occur in a neighboring position in a large corpus. Previous work by Hughes (1994) indicated that the two neighbors on the left and on the right (i.e. the words in positions $n-2$, $n-1$, $n+1$, $n+2$, relative to word $n$) are a good choice of context. The position of a word in Lexical Space is thus given by a four component vector, of which each component has as many dimensions as there are context-words. The value of a dimension is obtained from the frequency of occurrence of that context-word in that position. The continuous nature of the representations of words in Lexical Space allows us to represent graded categories, hierarchies, and categories of various granularity. The degree in which distributional patterns of words are informative of their syntactic category can now be quantified in the model. If distribution, which determines the structure of the Lexical Space, is important, then similarity (i.e. distance) of the distributions should map to syntactic similarity. In other words, are words that are represented in neighboring regions of Lexical Space members of the same category?

### 2.2 Experiments

To assess its utility we tested the Lexical Space approach on a large corpus of natural language. We used 3 million words of Wall Street Journal text, taken from the ACL/DCI CD-ROM I (1991). The corpus was available both as raw text, and in an annotated format where each word is tagged with a part of speech tag from the Penn Treebank Tagset (Marcus, Santorini & Marcinkiewicz 1993). This allows a direct comparison of the induced categories in Lexical Space with "human-assigned" categories. The possibility of evaluation outweighted arguments to use more realistic child-directed speech as input. For distributional experiments with parent-child discourse as input see e.g. Redington et al.(1995).

For the focus-words, i.e. the lexicon, we chose the 5000 most frequent words in the corpus. For the contexts, i.e. the dimensions of Lexical Space we experimented with the 250 and 1000 most frequent words. The distance metric used to measure the similarity between the normalized vectors is the dot-product. The vectors were normalized to unit length to discount for the absolute frequencies of the focus-words. The context-words also have a wide range of frequencies; the most frequent context words (mostly function words such as: the, of, as, that, in, it, etc.) get a very large share in the total distance between the vectors. To see what effect this has on the relatedness of words, we also used vectors in which each dimension is divided by the total frequency of that context. This enables low-frequent contexts to play a role as well, as long as they are more informative than chance. Without this manipulation there is little difference between the structure of the Lexical Space with just a few frequent context-words and one with a large amount of different contexts. The two types of normalization will be referred to below as $norm1$ and $norm2$ respectively.

### 2.3 Evaluation

Aiming to quantify how informative the distributional patterns in the corpus are, rather than just to rely on an intuitive evaluation and make the claim *that* they are informative,

we developed a method for comparing the structure of Lexical Space with the structure of a linguistic categorization scheme. For this purpose we assigned each of the 5000 focus-words its most frequent POS-tag in the annotated version of the corpus. The evaluation scheme measures for each word whether its neighbors are of the same category. The score is based on recall and precision. Because different categories have quite different numbers of members (see Table 2), it is not feasible to set a fixed neighborhood size for a category. As the neighborhood around a word under evaluation is increased, more and more words are included in it, some of the same category—which increases recall, some of another category—which lowers the precision (see Figure 1.). For each item in the neighborhood, its precision is the

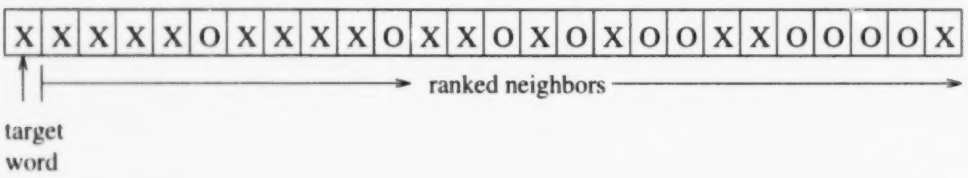


Figure 1: *Evaluation. The neighbors of a word ranked by their distance from the target word. X's are of the same category as the target (correct), O's are incorrect. The neighborhood is increased until all X's are in the list. Then the average precision (P) is computed over all X's.*

number of correct words in the neighborhood divided by the total number of words in the neighborhood. We define the measure of the quality of a word's neighborhood as the average precision of the correct words at 100% recall. This is called a word's P-value.

$$P = \frac{1}{N_c} \sum_{i \in correct} \frac{n_{ci}}{n_{retrieved\ i}}$$

Where *correct* is the set of words with the same category as the target word, $N_c$ is the number of words in this set and $n_{ci}$ and $n_{retrieved\ i}$ are respectively the number of correct words and the total number of words between correct word $i$ and the target word. Intuitively speaking, this measures the density of correct words in a neighborhood which includes the furthest correct word. If all correct words are closer than the closest incorrect word (best case) we get a P-value of 1, if all incorrect words are closer than the closest correct word (worst case) we get a P-value of close to zero. This measure has a number of problems: its worst case value is dependent on the size of the lexicon, and it does not deal with ambiguity (a word which is 60% noun, 40% verb in the corpus is taken as an incorrect neighbor of a word that is 40% noun, 60% verb, although in fact these words are quite similar in their behavior). The measure also assumes that the human-assigned categories are the ultimate standard of correct categorization, whereas in reality the correctness of one classification over another largely depends on the linguistic theory that is embraced. Nonetheless, we think that the measure is useful for the comparison of different words and categories, under various parameter settings of the Lexical Space model. Alternative measures of correlation between distributional and human-assigned categories can be found in Hughes & Attwell (1994) and in Redington et al. (1995). These measures, however, require a discrete clustering of the Lexical Space, which introduces additional inaccuracy.

### 3. Results

The structure of the resulting Lexical Space seems to confirm the expectations. A number of examples from different categories can be seen in Table 1 (at the end of this paper).

Direct neighbors of words belong to the same category, and fine grained distinctions can be made (e.g. within the noun region there are subregions for proper nouns, which are in turn made up of small neighborhoods, e.g. country names). This illustrates the fact that, because semantically similar words occur in similar linguistic contexts, some semantic distinctions are revealed in the distributional patterns, Another interesting observation can be made about the place of the words "should" and "ought" in the Lexical Space. While both are tagged as modal verbs in the Penn Treebank tagset, "should" (P: 0.852) clusters much nicer with the other modal verbs than "ought" (P: 0.175). "Ought" is much more similar to verbs which take complement arguments. This shows that in some cases the distributionally induced categories are much better at revealing certain fine lexical distinctions than the Penn Treebank tagset. It should be noted that in the quantitative evaluation these kind of categorization differences are disadvantegeous for the Lexical Space method.

To see whether similarity relations at a larger scale are also reflected in the overall structure of the Lexical Space we have used Principal Component Analysis (PCA) to display the Lexical Space in two dimensions. PCA identifies orthogonal vectors which cover the directions with the largest variance in the data (principal components). By projecting the original high-dimensional space onto a much lower number of principal components one obtains a large dimension reduction with maximum possible preservation of the distances between the points. The result of mapping a $4 \times 250$ dimensional Lexical Space, containing the 250 most frequent words, onto its first two principle components is shown in Figure 2.

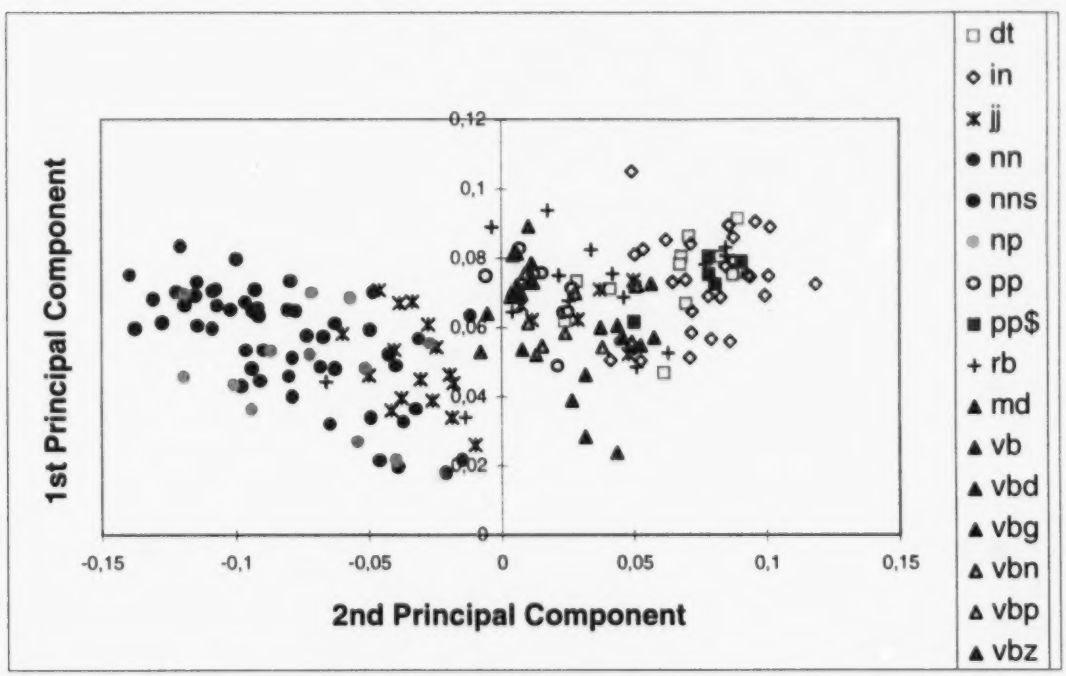


Figure 2: *Projection of a Lexical Space onto its first two Principal Components (lexicon:250, context:250, norm:1). determ.=dt, prep.=in, adj.=jj, nouns={nn,nns,np}, pers.pronoun(poss.)=pp($), adverb=rb, verbs={md, vb–vbz}*

One can see that in this projection the nouns are separated from the verbs quite well, and

that the adjectives lie somewhere in between the two. This is in line with a recent linguistic theory about the relation between these three categories (Wetzer 1995).

We think that from this quick glance at the structure of the categorization in Lexical Space one can safely conclude that Pinker's statement about the irrelevance of distributional patterns for bootstrapping is misguided (see also similar results by the authors cited in section 2). Moreover, the high quality and the level of detail of the groupings in the Lexical Space raise some questions about the status of distributional information as means of merely *bootstrapping* syntactic categories. We will return to this point in the discussion. We will now turn to a presentation of the quantitative data that we have compiled.

### 3.1 Quantitative evaluation

The analysis shows that not only the main open word-classes can be bootstrapped distributionally. In fact, all of the categories in the Penn Treebank tagset produce quite acceptable results. Nonetheless, there are differences between words and between categories as to how well their distributional behavior correlates with their syntactic properties. In the Lexical Space some words lie surrounded by neighbors with the same category, while others are less typical of their category and lie in their own separate region of the space, surrounded perhaps by a few similar exceptions. The P-value of words is indicative of such differences in typicality. This property of the representations in Lexical Space could possibly make the method a useful research tool for the theoretical study of syntax. We will now focus on the differences in P-value between different categories, to examine the limitations of the distributional information.

Table 2 shows the results of the evaluation averaged per category. Based on their average P-value, one can say that some classes are distributionally easier than others. That is, across all the experiments, the score of a category relative to other categories does not change much. Most large categories reach a P-value of over 0.4 which is quite good considering the fact that the larger classes of Penn Treebank tags really cover a number of smaller (sub)categories which are sometimes not very similar in distribution. Some of the categories, like cardinal numbers (*cd*), modal verbs (*md*), possesive pronouns (*pp*$) and verb infinitives (*vb*), cluster together really well in Lexical Space, reaching a P-value of around 0.6. On the other side there are some categories which perform poorly, most notably coordinating conjunctions (*cc*), adjectives (*jj*), plural proper nouns (*nps*), adverbs (*rb*) and the various verb inflections (*vbd*, *vbg*, *vbn*, *vbp*, *vbz*). These categories reveal some of the limitations of the Lexical Space combined with the limitations of the evaluation method.

The reason that the *cc* and *rb* categories are evaluating poorly is the fact that conjunctions and adverbs are distributionally relatively free. Adjectives are much less free, but in the context of the Wall Street Journal they are sometimes very similar to nouns, because many nouns in the corpus also function as modifiers of other nouns. An interesting question is whether "motor" in "Ford motor company" shouldn't really be viewed as something close to an adjective. Plural proper nouns receive that tag (*nps*) in the corpus because they are plural nouns that frequently occur in the all too often present company names. E.g. in "American Airlines" the word "Airlines" is considered a plural proper noun. In the Lexical Space it clusters with plural common nouns (*nns*), a logical choice, but one leading to a worse evaluation. Finally, a word remains to be said about the bad performance of the inflected verb forms. On detailed examination of the results it appeared that they are actually well separated from other main categories. However, they mingle with each other (e.g. present tense (*vbp*, *vbz*) and past tense (*vbd*) are distinguished in the tagset, but are distributionally almost indistinguishable). It seems that morphological information can be used to tease these subcategories apart. Another problem that reveals itself in the categorization of the verbs, is the ambiguity of words. In its present incarnation Lexical Space does not distinguish

between homonyms (but see e.g. Schütze (1995)).

In the experiments with $norm1$, the role a context dimension can play is limited by its frequency. With $norm2$ a context is in effect weighted by its informativeness. Therefore, as one can see in Table 2, the effect of increasing the number of context words is negligible under $norm1$. $Norm2$ has a worse score overall (0.37) than $norm1$ (0.43), and the quality of the grouping in Lexical Space gets even worse (0.31) when the number of contexts is increased to 1000. This shows that, as Pinker has already pointed out, the large amounts of possible distributional patterns can indeed make the bootstrapping task more difficult. However, we further investigated the possibility that the $norm2$ vectors do contain the same (or more) information as the $norm1$ vectors, but they are in some sense corrupted by noisy information. For this purpose we performed a PCA of the Lexical Space, compressing the vectors of the 250-experiment from 1000 to 15 dimensions, and those of the 1000-experiment from 4000 to 15 dimensions. The compression operation has the effect that only the most important regularities are kept, and that noise is discarded. After this treatment $norm1$ scored 0.46 and 0.45 for respectively 250 and 1000 sized contexts, and $norm2$ scored 0.48 and 0.43. This shows that Pinker's efficiency problem might not be so hard to deal with after all.

## 4. Discussion

With the aid of the Lexical Space method we have analyzed the distributional information available in the linguistic environment, in casu a large corpus of text. This excercise has shown that the discussion between Maratsos and Pinker can be settled by empirical evidence. The co-occurrence and adjacency information of words is highly predictive of their syntactic category, because in the Lexical Space words in each other's neighborhood tend to have the same syntactic category. The experiments with different numbers of context features also indicate that within the Lexical Space model the very high number of (possibly contradictory) features poses no serious problems for bootstrapping. The analysis that we have presented did not explicitly go into two related questions. The first of these is whether the available distributional information can be enough to bootstrap syntactic categories. We think that children use a varied blend of information sources, although distributional information has been the easiest to study using the Lexical Space method so far. We do wish to point out, however, that other information sources can easily be incorporated as extra dimensions into the Lexical Space. One might e.g. add the inflection morpheme of a word as a co-occurring context. The claim that a certain combination of information sources is "enough" for bootstrapping into syntax, really is too dependent on what one exactly means by syntax or syntactic categories. The second question is whether the child indeed uses the available information, and if so, which mechanisms it employs for this purpose.

So far we have viewed the Lexical Space as a method for the analysis of the language environment. Another outlook on the Lexical Space is to take it more or less seriously as a model of the child's representations and processes. The quality and the level of detail with which distinctions are made in the vector representations certainly do suggest that such a view might be advantageous. An attractive feature of such a position is that it offers a more empiricist view on lexical syntactic categories. In Lexical Space categories emerge from the patterns in the data so that they do not have to be assumed to be innate. In the space syntactic categories of words are represented by their position. Words of the same or similar category are represented in nearby regions. The Lexical Space allows a continuous, fine-grained representation of lexical categories, so that a lexical specification of many syntactic constraints seems possible (MacDonald, Pearlmutter & Seidenberg 1994). In the Lexical Space sequences of words are in fact trajectories through the space.

This opens up a new way of studying the acquisition of syntax in terms of pattern recognition of temporal sequences. Veenstra & Zavrel (1994) give some ideas how we can interpret grammatical sentences as trajectories through Lexical Space.

## References


Bates, E. & B. MacWhinney (1987): "Competition, variation, and language learning." In: B. MacWhinney (ed.) *Mechanisms of Language Acquisition*. Lawrence Erlbaum Associates, Hillsdale, NJ

Braine, M. (1992): "What sort of innate structure is needed to "bootstrap" into syntax?." *Cognition* 45:77–100

Finch, S. P. (1993): *Finding Structure in Language*. PhD. Thesis, Centre for Cognitive Science, University of Edinburgh

Hughes, J. (1994): *Automatically Acquiring a Classification of Words*. PhD. Thesis, School of Computer Studies, University of Leeds

Hughes, J. & E. Attwell (1994): "The automated evaluation of inferred word classification." In: A. Cohn (ed.) *Proc. of the 11th European Conference on Artificial Intelligence (ECAI94)*. John Wiley & Sons, Ltd., pp. 535–539

MacDonald, M. C., N. J. Pearlmutter & M. S. Seidenberg (1994): "Lexical nature of syntactic ambiguity resolution." *Psychological Review* 101: 4:676–703

Maratsos, M. P. (1982): "The child's construction of grammatical categories." In: E. Wanner & L. R. Gleitman (eds.) *Language Acquisition: the state of the art*. Cambridge University Press

Maratsos, M. P. (1990): "Are actions to verbs as objects are to nouns? On the differential semantic bases of form, class, category." *Linguistics* 28:1351–1379

Maratsos, M. P. & M. Chalkley (1981): "The internal language of children's syntax: the ontogenesis and representation of syntactic categories." In: K. Nelson (ed.) *Children's Language*. Gardner Press, New York

Marcus, M. P., B. Santorini & M. A. Marcinkiewicz (1993): "Building a large annotated corpus of English: The Penn Treebank." *Computational Linguistics* 19:313–330

Pinker, S. (1984): *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA

Redington, M., N. Chater & S. Finch (1995): *The potential contribution of distributional information to early syntactic category acquisition*. submitted paper

Schütze, H. (1995): *Ambiguity in Language Learning: Computational and Cognitive Models*. PhD. Thesis, Department of Linguistics, Stanford University

Veenstra, J. B. & J. Zavrel (1994): "A corpus-based approach to syntax acquisition." *Proc. of the 3rd conference on the Cognitive Science of Natural Language Processing*, Dublin City University

Wetzer, H. (1995): *Nouniness and verbiness: A typological study of adjectival predication*. PhD. Thesis, Faculty of Arts, University of Nijmegen




BEST COPY AVAILABLE

| | | | | | |
|---|---|---|---|---|---|
| IN | *in:54426* | rb:158 | rp:269 | **P: 0.680** | |
| for(in)0.05 | since(in)0.10 | at(in)0.11 | after(in)0.11 | under(in)0.11 | on(in)0.12 |
| until(in)0.12 | by(in)0.13 | among(in)0.14 | before(in)0.16 | without(in)0.17 | against(in)0.19 |
| GROUP | *nn:2069* | np:962 | **P: 0.678** | | |
| network(nn)0.08 | firm(nn)0.11 | measure(nn)0.11 | package(nn)0.11 | chain(nn)0.11 | club(np)0.11 |
| bill(nn)0.11 | partnership(nn)0.12 | panel(nn)0.12 | fund(nn)0.12 | team(nn)0.12 | party(nn)0.13 |
| JAPAN | *np:1245* | **P: 0.423** | | | |
| china(np)0.16 | france(np)0.16 | britain(np)0.19 | canada(np)0.19 | mexico(np)0.19 | india(np)0.19 |
| australia(np)0.20 | korea(np)0.22 | italy(np)0.23 | detroit(np)0.23 | greece(np)0.24 | africa(np)0.24 |
| HE | *pp:11424* | **P: 0.189** | | | |
| she(pp)0.032 | traders(nns)0.20 | analysts(nns)0.29 | sources(nns)0.29 | gillette(np)0.30 | at&t(np)0.32 |
| it(pp)0.32 | dealers(nns)0.32 | they(pp)0.33 | ibm(np)0.33 | nasa(np)0.33 | officials(nns)0.34 |
| HIS | *pp$:5729* | **P: 1.00** | | | |
| our(pp$)0.13 | its(pp$)0.13 | their(pp$)0.14 | my(pp$)0.16 | her(pp$)0.18 | your(pp$)0.18 |
| the(dt)0.23 | another(dt)0.33 | every(dt)0.34 | those(dt)0.35 | a(dt)0.38 | columbia(np)0.38 |
| SHOULD | *md:1354* | **P: 0.852** | | | |
| must(md)0.07 | could(md)0.08 | can(md)0.08 | may(md)0.09 | would(md)0.09 | might(md)0.09 |
| will(md)0.11 | 'll(md)0.24 | had(vbd)0.35 | wo(md)0.36 | has(vbz)0.36 | is(vbz)0.37 |
| OUGHT | *md:77* | **P: 0.175** | | | |
| tend(vbp)0.17 | seems(vbz)0.24 | appears(vbz)0.25 | appeared(vbd)0.25 | seemed(vbd)0.27 | continues(vbz)0.27 |
| need(vbp)0.34 | promises(vbz)0.35 | seem(vbp)0.35 | thought(vbd)0.35 | want(vbp)0.35 | wants(vbz)0.35 |
| DISMISSED | *vbd:44* | vbn:31 | **P: 0.210** | | |
| rejected(vbd)0.31 | described(vbd)0.33 | cited(vbd)0.34 | triggered(vbd)0.34 | hit(vbd)0.34 | fired(vbn)0.35 |
| given(vbn)0.36 | discovered(vbn)0.37 | passed(vbn)0.37 | made(vbn)0.37 | brought(vbn)0.38 | approved(vbd)0.38 |

Table 1: *Some examples of the direct neighbors of words in a Lexical Space (context:250 lexicon:5000 norm:1). The word in uppercase is the word under evaluation, and the line after the word lists the category frequencies in the annotated corpus and the overall P-value. The lines after that list the 12 nearest neighbors of the target word, ranked by ascending distance.*

| Tag:No. | 250.1 | 250.2 | 1000.1 | 1000.2 | Tag:No. | 250.1 | 250.2 | 1000.1 | 1000.2 |
|---|---|---|---|---|---|---|---|---|---|
| $ : 5 | 0.92 | 0.75 | 0.78 | 0.75 | pos : 2 | 0.54 | 0.81 | 0.61 | 0.72 |
| . : 2 | 0.83 | 0.85 | 0.83 | 0.8 | pp : 16 | 0.35 | 0.38 | 0.36 | 0.42 |
| : : 4 | 0.34 | 0.4 | 0.34 | 0.32 | pp$ : 7 | 1 | 0.94 | 0.99 | 0.82 |
| cc : 8 | 0.22 | 0.16 | 0.24 | 0.18 | rb : 187 | 0.17 | 0.13 | 0.17 | 0.12 |
| cd : 228 | 0.58 | 0.68 | 0.58 | 0.66 | rbr : 3 | 0.34 | 0.35 | 0.34 | 0.35 |
| dt : 14 | 0.42 | 0.5 | 0.46 | 0.48 | rp : 2 | 0.92 | 0.92 | 0.92 | 0.92 |
| in : 57 | 0.47 | 0.34 | 0.46 | 0.2 | vb : 255 | 0.67 | 0.61 | 0.67 | 0.55 |
| jj : 524 | 0.26 | 0.2 | 0.27 | 0.2 | vbd : 188 | 0.29 | 0.2 | 0.29 | 0.18 |
| jjr : 21 | 0.37 | 0.46 | 0.36 | 0.39 | vbg : 203 | 0.2 | 0.18 | 0.19 | 0.17 |
| jjs : 11 | 0.29 | 0.33 | 0.28 | 0.34 | vbn : 267 | 0.34 | 0.41 | 0.34 | 0.28 |
| md : 13 | 0.68 | 0.74 | 0.68 | 0.65 | vbp : 33 | 0.25 | 0.23 | 0.26 | 0.24 |
| nn : 1301 | 0.58 | 0.41 | 0.58 | 0.34 | vbz : 110 | 0.27 | 0.24 | 0.26 | 0.13 |
| nns : 656 | 0.42 | 0.4 | 0.41 | 0.29 | wdt : 2 | 0.5 | 0.5 | 0.5 | 0.5 |
| np : 845 | 0.42 | 0.34 | 0.39 | 0.28 | wp : 3 | 0.37 | 0.4 | 0.38 | 0.59 |
| nps : 18 | 0.11 | 0.11 | 0.12 | 0.096 | wrb : 4 | 0.43 | 0.43 | 0.45 | 0.43 |
| | | | | | Totals:5000 | 0.43 | 0.37 | 0.43 | 0.31 |

Table 2: *Per category evaluation at various settings of the Lexical Space model. The rows in this table give the P-value averaged over all words in a category (given in the first collumn; the number in the first collumn gives the total number of word types in that class). The next four collumns stand for four different experiments. The size of the lexicon is 5000 words across all experiments. The number of context words varies between 250 and 1000, and the normalization varies between 1 and 2. The averages over all words for the four experiments are in the bottom row.*



BEST COPY AVAILABLE